CMPT 409/981: Optimization for Machine Learning

Lecture 16

Sharan Vaswani

November 5, 2024

Recap

Generic Online Optimization (w_0, Algorithm $\mathcal{A}$, Convex set $\mathcal{C} \subseteq \mathbb{R}^d$)

for $k = 1, \ldots, T$ **do**
  Algorithm $\mathcal{A}$ chooses point (decision) $w_k \in \mathcal{C}$
  Environment chooses and reveals the (potentially adversarial) loss function $f_k : \mathcal{C} \to \mathbb{R}$
  Algorithm suffers a cost $f_k(w_k)$
end for

- **Regret**: For any fixed decision $u \in \mathcal{C}$, $R_T(u) := \sum_{k=1}^{T} [f_k(w_k) - f_k(u)]$.

- **Online Gradient Descent** (OGD): At iteration k, the algorithm chooses the point w_k. After the loss function f_k is revealed, OGD suffers a cost $f_k(w_k)$ and uses the function to compute: $w_{k+1} = \Pi_C[w_k - \eta_k \nabla f_k(w_k)]$ where $\Pi_C[x] = \arg\min_{y \in \mathcal{C}} \frac{1}{2} \|y - x\|^2$.

- **Claim**: If the convex set $\mathcal{C}$ has a diameter D i.e. for all $x, y \in \mathcal{C}$, $\|x - y\| \leq D$, for an arbitrary sequence of losses such that each f_k is convex, differentiable and G-Lipschitz, OGD with $\eta_k = \frac{\eta}{\sqrt{k}}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$, $R_T(u) \leq \frac{D^2 \sqrt{T}}{2\eta} + G^2 \sqrt{T}\, \eta$.

Recap

- Given a differentiable, strictly-convex mirror map ϕ, $D_\phi(y,x) := \phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle$.
- **Online Mirror Descent** (OMD): $w_{k+1} = \arg\min_{w \in \mathcal{C}} \left[\langle \nabla f_k(w_k), w \rangle + \frac{1}{\eta_k} D_\phi(w, w_k) \right]$.

Setting $\phi(x) = \frac{1}{2} \|x\|^2$ results in $D_\phi(y,x) = \frac{1}{2} \|y - x\|^2$ and recovers OGD.

- *Example*: For prediction with expert advice, $\mathcal{C} = \Delta_d = \{w_i | w_i \geq 0 \; ; \; \sum_{i=1}^d w_i = 1\}$ and we typically use the *negative-entropy mirror map* i.e. $\phi(w) = \sum_{i=1}^d w_i \ln(w_i)$. In this case, $D_\phi(u,v) = \text{KL}(u||v)$.
- The OMD update can be equivalently written as:

GD in dual space: $w_{k+1/2} = (\nabla \phi)^{-1} \left(\nabla \phi(w_k) - \eta_k \nabla f_k(w_k) \right)$
Bregman projection: $w_{k+1} = \arg\min_{w \in \mathcal{C}} D_\phi(w, w_{k+1/2})$

- With the negative-entropy mirror map, OMD results in the **multiplicative weights update**:

$w_{k+1}[i] = \frac{w_k[i] \, \exp(-\eta_k \, g_k[i])}{\sum_{j=1}^d w_k[j] \, \exp(-\eta_k \, g_k[j])}$.

Online Mirror Descent – Convex, Lipschitz functions

In order to analyze OMD, we will make some assumptions about $\mathcal{C}$, f_k and ϕ.

- **Assumption 1**: $\mathcal{C}$ is a convex set and $\forall k$, f_k is a convex function.
- **Assumption 2**: $\forall k$, f_k is G-Lipschitz in the ℓ_p norm (for $p \geq 1$), implying that $\forall w \in \mathcal{C}$,

$$\|\nabla f_k(w)\|_p \leq G$$

- **Assumption 3**: ϕ is ν strongly-convex in the ℓ_q norm (for $q \geq 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$) i.e.

$$\phi(y) \geq \phi(x) + \langle \nabla \phi(x), y - x \rangle + \frac{\nu}{2} \|y - x\|_q^2$$

Example: For prediction from expert advice,

- $\mathcal{C} = \Delta_d$ is a convex set and $f_k(w_k) = \langle c_k, w_k \rangle$ is a convex function.
- If the costs are bounded by M, then, $\|\nabla f_k(w)\|_\infty = \|c_k\|_\infty \leq M$. Hence, $p = \infty$, $G = M$.
- If $\phi(w)$ is negative-entropy, then by Pinsker's inequality, $q = 1$ and $\nu = 1$ i.e.

$$\phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle = D_\phi(y, x) = \mathrm{KL}(y||x) \geq \frac{1}{2} \|y - x\|_1^2 .$$

Online Mirror Descent – Convex, Lipschitz functions

Claim: For an arbitrary sequence of losses such that each f_k is convex, G-Lipschitz and differentiable, then OMD with a ν strongly-convex mirror map ϕ, $\eta_k = \eta = \sqrt{\frac{2\nu}{T}} \frac{D}{G}$ where $D^2 := \max_{u \in \mathcal{C}} D_\phi(u, w_1)$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{\sqrt{2}\, DG}{\sqrt{\nu}} \sqrt{T}\,,$$

Proof: Recall the mirror descent update: $\nabla\phi(w_{k+1/2}) = \nabla\phi(w_k) - \eta_k \nabla f_k(w_k)$. Setting $\eta_k = \eta$ and using the definition of regret,

$$\begin{aligned} R_T(u) &= \sum_{k=1}^{T} f_k(w_k) - f_k(u) \leq \sum_{k=1}^{T} [\langle g_k, w_k - u \rangle] && \text{(Convexity of } f_k \text{ and } g_k := \nabla f_k(w_k)) \\ &= \sum_{k=1}^{T} \frac{1}{\eta} \left\langle \nabla\phi(w_k) - \nabla\phi(w_{k+1/2}), w_k - u \right\rangle && \text{(Using the OMD update)} \end{aligned}$$

Online Mirror Descent – Convex, Lipschitz functions

Recall that $R_T(u) = \sum_{k=1}^{T} \frac{1}{\eta} \left\langle \nabla\phi(w_k) - \nabla\phi(w_{k+1/2}), w_k - u \right\rangle$

Three point property: for any 3 points x, y, z,

$$\langle \nabla\phi(z) - \nabla\phi(y), z - x\rangle = D_\phi(x, z) + D_\phi(z, y) - D_\phi(x, y)$$

$$\left\langle \nabla\phi(w_k) - \nabla\phi(w_{k+1/2}), w_k - u \right\rangle = D_\phi(u, w_k) + D_\phi(w_k, w_{k+1/2}) - D_\phi(u, w_{k+1/2})$$

$$\implies R_T(u) \leq \sum_{k=1}^{T} \frac{1}{\eta} \left[D_\phi(u, w_k) + D_\phi(w_k, w_{k+1/2}) - D_\phi(u, w_{k+1/2}) \right]$$

From the OMD update, we know that, $w_{k+1} = \arg\min_{w \in \mathcal{W}} D_\phi(w, w_{k+1/2})$. Recall the optimality condition: for a convex function f and a convex set $\mathcal{C}$, if $x^* = \arg\min_{x \in \mathcal{C}} f(x)$, then $\forall x \in \mathcal{X}$, $\langle \nabla f(x^*), x^* - x\rangle \leq 0$. Using this condition for $D_\phi(w, w_{k+1/2})$, for $u \in \mathcal{C}$,

$$\left\langle \nabla\phi(w_{k+1}) - \nabla\phi(w_{k+1/2}), w_{k+1} - u \right\rangle \leq 0$$

$$\implies -D_\phi(u, w_{k+1/2}) \leq -D_\phi(u, w_{k+1}) - D_\phi(w_{k+1}, w_{k+1/2}) \qquad \text{(3 point property)}$$

$$\implies R_T(u) \leq \sum_{k=1}^{T} \frac{1}{\eta} \left[D_\phi(u, w_k) - D_\phi(u, w_{k+1}) \right] + \frac{1}{\eta} \left[D_\phi(w_k, w_{k+1/2}) - D_\phi(w_{k+1}, w_{k+1/2}) \right]$$

Online Mirror Descent – Convex, Lipschitz functions

Telescoping we conclude that $R_T(u) \leq \frac{1}{\eta} D_\phi(u, w_1) + \frac{1}{\eta} \sum_{k=1}^{T} \left[D_\phi(w_k, w_{k+1/2}) - D_\phi(w_{k+1}, w_{k+1/2}) \right]$.

$$
\begin{aligned}
D_\phi(w_k, w_{k+1/2}) - D_\phi(w_{k+1}, w_{k+1/2}) &= \phi(w_k) - \phi(w_{k+1}) - \langle \nabla\phi(w_{k+1/2}), w_k - w_{k+1} \rangle \\
&\leq \langle \nabla\phi(w_k) - \nabla\phi(w_{k+1/2}), w_k - w_{k+1} \rangle - \frac{\nu}{2} \|w_k - w_{k+1}\|_q^2 \\
&\qquad \text{(Using strong-convexity of } \phi \text{ with } y = w_{k+1} \text{ and } x = w_k\text{)} \\
&= \eta \langle g_k, w_k - w_{k+1} \rangle - \frac{\nu}{2} \|w_k - w_{k+1}\|_q^2 \qquad \text{(Using the OMD update)} \\
&\leq \eta G \|w_k - w_{k+1}\|_q - \frac{\nu}{2} \|w_k - w_{k+1}\|_q^2 \\
&\qquad \text{(Holder's inequality: } \langle x, y \rangle \leq \|x\|_p \|y\|_q \text{ s.t. } \tfrac{1}{p} + \tfrac{1}{q} = 1 \text{ and since } \|g_k\|_p \leq G\text{)} \\
&\leq \frac{\eta^2 G^2}{2\nu} \qquad \text{(For all } z,\ a\,z - bz^2 \leq \tfrac{a^2}{4b}\text{)}
\end{aligned}
$$

$$
\implies R_T(u) \leq \frac{1}{\eta} D_\phi(u, w_1) + \frac{\eta G^2\, T}{2\nu} \leq \frac{D^2}{\eta} + \frac{\eta G^2\, T}{2\nu} \qquad \text{(Since } D_\phi(u, w_1) \leq D^2\text{)}
$$

$$
\implies R_T(u) \leq \frac{\sqrt{2} D G}{\sqrt{\nu}} \sqrt{T} \qquad \text{(Setting } \eta = \sqrt{\tfrac{2\nu}{T}}\, \tfrac{D}{G}\text{)}
$$

Online Mirror Descent – Example

We have proved that for any fixed comparator u, $R_T(u) \leq \frac{\sqrt{2}DG}{\sqrt{\nu}} \sqrt{T}$ where,
(i) $\|\nabla f_k(w)\|_p \leq G$, (ii) $\phi(y) \geq \phi(x) + \langle \nabla \phi(x), y - x \rangle + \frac{\nu}{2} \|y - x\|_q^2$ and (iii) $D_\phi(u, w_1) \leq D^2$.

• Using OMD with negative-entropy for prediction with expert advice, $p = \infty$, $q = 1$, $\nu = 1$. Since $\|c_k\|_\infty \leq M$, $G = M$. If $\forall i \in [d]$, $w_1[i] = \frac{1}{d}$, $D_\phi(u, w_1) = \sum_{i=1}^{d} u_i \ln(u_i\, d) \leq \ln(d)$.

$$\implies R_T(u) \leq \sqrt{2}M \sqrt{\ln(d)} \sqrt{T}$$

• Since OGD is a special case of OMD with $\phi(w) = \frac{1}{2} \|w\|^2$, using OGD for prediction with expert advice, $p = 2$, $q = 2$, $\nu = 1$. Since $\|c_k\|_\infty \leq M$, using the relation between norms, $G = M\sqrt{d}$. If $\forall i \in [d]$, $w_1[i] = \frac{1}{d}$, $D_\phi(u, w_1) = \frac{1}{2} \|u - w_1\|^2 \leq \sqrt{2}$

$$\implies R_T(u) \leq 2M \sqrt{d} \sqrt{T}$$

• Hence, using multiplicative weights results in $O(\sqrt{\ln(d)}\sqrt{T})$ regret which is better than the $O(\sqrt{d}\sqrt{T})$ regret obtained by OGD. For prediction with expert advice, when the number of experts is large, this can be a substantial advantage.

Questions?